ISSN 0317-2473

Research Discussion Paper No. 69

LOG-LINEAR MODELLING AND SIMULTANEOUS STANDARDIZATION OF RATES

By

Frank Trovato

Department of Sociology
University of Alberta
Edmonton, Alberta

November 1990

92-06285/

ABSTRACT

Social scientists are often interested in analyzing rates for rare events such as homicide, suicide, and so on. This paper demonstrates how rates can be modelled and standardized for age and gender compositions simultaneously using log-linear rate models. SPSSx LOGLINEAR can be applied to achieve this task by using the exposures for the cell counts as weights. The parameter estimates from fitted models are easily interpretable as the degree to which variables add to or subtract from the overall rate. The log-linear rate model with simultaneous standardization is a flexible method to test competing hypotheses. Youth suicide in Canada for 1971 and 1981 is used as an empirical example to demonstrate the application of log-linear rate modelling with simultaneous standardization.

1. Introduction

Social scientists are often interested in analyzing rates pertaining to rare phenomena of interest to their disciplines. For example, one may be interested in uncovering the factors associated with changes in the homicide or suicide rate over some period of time. A popular approach developed by demographers to analyze rates is the method of direct or indirect standardization (Shryock and Siegel, 1973; Kitagawa, 1955; 1964; Clogg and Eliason, 1987; 1988). Lalu (1987) has given an exposition of these standardization methods and their application to criminology in an earlier issue of this journal.

In this paper it is demonstrated how log-linear models can be adapted to perform standardization of rates. The first part of the paper develops the basic principles inherent in log-linear rate models and standardization. It is assumed that the reader is familiar with log-linear models. Those not familiar with this methodology may consult Haberman (1978; 1979) and Clogg and Eliason (1987). The second part of the article applies the methods developed in the initial section to an empirical example, that of youth suicide in Canada for the periods 1971 and 1981.

2. Modelling Log of Rates

The method of fitting models to log of rates (called hereafter as log-rate models) is an extension of the general log-linear model for countable data (Clogg and Eliason, 1987). In the latter case, the interest is usually in fitting alternate models (equations) to frequency data to find one that provides a good fit to the data. A good model is one whose parameters reproduce almost perfectly the observed marginals for a given crosstabulation. For example, in the case of two variables, A and B, the following saturated model for the cell counts, f_{ij}, will reproduce the data perfectly:

$$\log(f_{ij}) = \lambda + \lambda^{A}_{i} + \lambda^{B}_{j} + \lambda^{AB}_{ij}$$

$$i = 1,\ldots,I \quad j = 1,\ldots,J,$$

where λ is an overall parameter, λ^{A}_{i} the parameter for variable A, λ^{B}_{j} the parameter for variable B, and λ^{AB}_{ij} the interaction term of these two variables. In this case, the expected marginals for A and B will be equal to the observed marginals for these variables.

In most applied situations, analysts are interested in finding a model that is parsimonious, in that it describes the data with as few interaction terms as possible. Thus, in our example, an alternative model might be:

$$\log (f_{ij}) = \lambda + \lambda^{A}_{i} + \lambda^{B}_{j} .$$

Computer packages such as BMDP, SAS and SPSSx can be used to fit log-linear models. They employ maximum likelihood algorithms to estimate parameters, such as iterative proportional fitting (IPF), or the Newton-Raphson algorithm, used in SPSSx.

Clogg and Eliason (1987) have mentioned that the log-linear model can be easily extended to simultaneously standardize and model rates in the case where the dependent variable is dichotomous. For example, assume that we wish to model the suicide rate for male and female youths aged 10 to 29 in 1971 and 1981. An appropriate model may be specified as follows:

$$\log(d_{ijk}/E_{ijk}) = \lambda + \lambda^{A}_{i} + \lambda^{S}_{j} + \lambda^{Y}_{k} + \lambda^{AS}_{ij},$$

$$i=1,\ldots,I,\ j=1,2,\ K=1,2,$$

with constraints $\Sigma\lambda^{A}_{i} = \Sigma\lambda^{S}_{j} = \Sigma\lambda^{Y}_{k} = \Sigma\lambda^{AS}_{ij} = 0,$

where d_{ijk} = suicide counts by age, gender and year;

E_{ijk} = the population exposed to the risk of suicide by age, gender and year,

λ = intercept term,

λ^{A}_{i} = parameter for age effects,

λ_j^S = parameter for sex effects,

λ_k^Y = parameter for year effects,

λ_{ij}^{AS} = parameter for age-sex interaction effects.

Note that the left side of (3) is the natural logarithm of the suicide rate, whereas in (2), the left side is the natural logarithm of the cell frequencies (e.g., suicides). The above equation is expressed in additive terms, but can be transformed to be multiplicative by taking the antilogarithm of each term.

Many analysts are now familiar with fitting log-linear frequency models, but the adapting of the log-linear approach to fit rates has not received as much attention. Appendix A provides an example of fitting log-linear rate models to youth suicide rate in Canada using the SPSSx LOGLINEAR subroutine.

3. Direct Standardization with Log-Rate Models

Lalu (1987) has given a detailed account of both direct and indirect standardization, therefore only the rudimentary principles are considered here. For all practical purposes, the method of direct standardization is preferrable to the indirect one because it utilizes the full knowledge of age-specific rates in the study population. In the indirect method, the analyst does not

possess such information for the study group and is forced to use the age-specific rates of a "standard population" to make inferences about the group being analyzed.

Direct standardization removes the confounding effects of age composition from the observed rate of the study population. For example, in the case of youth suicide (ages 10-29), between 1971 and 1981 there has been an increase in the rate from 8.61 per 100,000 population in 1971, to 12.98 in 1981. The difference between these two rates is 4.37. The question to ask is how much of this difference is accounted by changes in the age composition of youths? All other things being equal, an increase in young people would account for an inrease in the rate of suicide. If we remove the confounding effects of age composition, we may find that the increase in suicide over time is perhaps less than 4.37 per 100,000 population.

The method of direct standardization may be executed by applying the age-specific suicide rates in 1981 to the age-specific population of 1971. This procedure would yield the expected number of suicides that would occur in 1981 if age composition remained constant. Thus, the age standardized suicide rate for 1981 would be:

$$SR^*_{81} = \frac{\Sigma(\frac{d_i}{E_i} \cdot PS_i)}{\Sigma PS_i} \quad , i = 1, \ldots, I$$

where SR^*_{81} = age-standardized suicide rate for 1981;

d_i = suicide deaths by age,

E_i = age-specific population in 1981,

PS_i = age-specific population in 1971 (the standard population).

This information can be incorporated in the log-rate model to control for age composition as follows:

$$\log ((d_{ik}/E_{ik}) \cdot (PS_i/ \Sigma PS_i)) = \lambda + \lambda_i^A + \lambda_k^Y + \lambda_{ik}^{AY}$$

where d_{ik} = suicide deaths by age and year,

E_{ik} = population exposed to the risk of suicide by age and year,

PS_i = 1971 age-specific standard population,

λ = an intercept term,

λ_i^A = age parameter,

λ_k^Y = year effect,

λ_{ik}^{AY} = age-year interaction effect.

This formula models the logarithm of the suicide rate while simultaneously standardizing for age composition. We may also standardize for sex composition since it is a well-documented fact that male suicides outnumber female suicides. Therefore, all other things being equal, an

increase in males over time would contribute to an increase in suicide. We can remove the sex compositional effect as follows:

$$\log\left((d_{ijk}/E_{ijk}).(PS_{ij}/\Sigma PS_{ij})\right) = \lambda + \lambda^{A}_{i} + \lambda^{S}_{j} + \lambda^{Y}_{k} + \lambda^{AS}_{ij} + \lambda^{AY}_{ik},$$

$$i = 1,..,I \quad j = 1,...,J \quad k = 1,...,K$$

The difference between (5) and (6) is the inclusion of terms for a sex effect λ^{S}_{j}, and an age-sex interaction term λ^{AS}_{ij}, where

d_{ijk} = suicide deaths by age, sex and year,

E_{ijk} = population exposed to the risk of suicide by age, sex and year,

PS_{ij} = age-sex specific standard population.

The remaining terms in (6) are as defined in (5).

Equation (6) models the effects of age, sex, year and age-sex interactions on the logarithm of the suicide rate net of age and sex compositional effects; that is, it assumes constant age and sex compositions over time.

The parameters in the equation are interpreted as the extent to which they raise or lower the overall suicide rate, net of age-sex composition and variables in the model. This interpretation is facilitated by the fact that in SPSSx LOGLINEAR the design matrix for categorical variables is specified in terms of -1's, + 1's. This fact

ensures the identification constraint of all terms in the equation summing to zero.

How are the age and sex parameters to be interpreted since they have been "purged" of their compositional effects? Obviously these two variables subsume many possible sources of variation that cannot be easily separated. Nevertheless, these two variables are net of their compositional effects in the log-rate model specified.

4. Empirical Example: Youth Suicide in Canada, 1971 and 1981

Table 1 shows data on youth suicide in Canada for 1971 and 1981. From this information we see that the male suicide rate surpasses that of females by a ratio of 4:1. The overall suicide rate increased by 51 percent between 1971 and 1981. The male rate increased by 56 percent, while that of females changed by 31 percent.

In conventional standardization one would compute the directly standardized rates for males and females separately for 1981 based on the 1971 population, or for both sexes combined, as done in Table 1. The overall standardized rate is 12.38, while the overall unstandardized rate is 12.98. The difference between these two rates is due to changes in age and sex compositions. Judging from the small difference, there seems to have been

TABLE 1

Suicides, Population Counts and Rates for Youths Aged 10-29 in Canada, 1971 and 1981

1971	Suicides		Population		Rates (per 100,000)		Total Age-Specific Rates
Age	Males	Females	Males	Females	Males	Females	
10-14	13	4	1181450	1129285	1.10	0.40	0.74
15-19	136	32	1074435	1039915	12.66	3.08	7.95
20-24	218	54	941775	947630	23.15	5.70	14.40
25-29	178	45	800710	783410	22.23	5.74	14.08
TOTAL	545	135	3998370	3900240	13.63	3.46	8.61
1981							
Age							
10-14	25	9	984740	936130	2.54	1.00	3.63
15-19	250	43	1182015	1132875	21.15	3.80	12.66
20-24	390	69	1174295	1169520	33.21	5.90	19.58
25-29	276	75	1084410	1093200	25.45	6.86	16.12
TOTAL	941	196	4425460	4331725	21.26	4.52	12.98
% Change 1971-81:	(72.7%)	(45.2%)	(10.7%)	(11.1%)	(56%)	(31%)	(51%)

NOTE: Total Suicide rate for both years = 10.91

1981 Age Standardized rate on the basis of 1971 Population = 12.38

Data Sources: Vital Statistics for 1971 and 1981. Census of the Population for 1971 and 1981.

little change in the effects of these compositions on suicide.

Table 2 displays eight log-rate models with simultaneous standardization. The model containing only age effects is used to judge the extent to which alternative models improve the fit to the suicide data in Table 1. The baseline log-likelihood chi-square (L^2) is 857.97 with 12 degrees of freedom, and the model does not fit the data very well. Adding sex and year to the equation reduces L^2 considerably, but not enough to provide a good fitting model. Model 4, which contains the main effects of age, sex and year, plus the interaction of age and year, provides a good fit to the data. All the other models were evaluated as possible alternatives, but they do not add significantly or parsimoniously to the information contained in equation number 4 in Table 2.

Table 3 shows predicted suicides, expected rates and standardized residuals obtained from model 4. The standardized residuals are all below 2.00, indicating that we have selected a good fitting model; that is, the differences between the observed and the predicted suicide counts are insignificant. Consequently, the predicted rates do not differ much from the observed rates in Table 1.

Table 4 contains the parameter estimates of model 4, in additive and multiplicative versions. The highest odds of suicide are associated with the age group 20-24, whereas

TABLE 2

Log-Rate Models Fitted to the Youth Suicide Data

Model	L^2	DF	P	Decomposition			
				Models Compared	ΔL^2	DF	P
1. A	857.97	12	.000	...			
2. A,S	61.89	11	.000	L_2-L_1	796.08	1	$\leq .01$
3. A,S,Y	19.89	10	.030	L_3-L_2	42.00	1	$\leq .01$
4. A,S,Y,AY	9.87	7	.196	L_4-L_3	10.02	3	$\leq .01$
5. A,S,Y,AS	14.15	7	.049	L_5-L_3	5.74	3	NS
6. A,S,Y,SY	17.88	9	.037	L_6-L_3	2.01	1	NS
7. A,S,Y,AY,AS	4.10	4	.392	L_7-L_4	5.77	3	NS
8. A,S,Y,AY,AS,SY	2.12	3	.547	L_8-L_4	7.75	4	NS

NOTE: A = Age (1=10-14, 2=15-19, 3=20-24, 4=25-29)
S = Sex (1 = Male, 2 = Female)
Y = Year (1 = 1971, 2 = 1981)

TABLE 3

Expected Suicides, Predicted Rates and Standardized Residuals Obtained with Model 4

1971	d_{ij}	r_{ij}	Standardized Residuals
Males			
10-14	13.91	1.18	-.2432
15-19	137.43	12.79	-.1224
20-24	222.51	23.63	-.3026
25-29	182.43	22.78	-.3279
Total	556.28	13.91	
Females			
10-14	3.09	0.27	.5158
15-19	30.57	2.94	.2596
20-24	49.49	5.22	.6416
25-29	40.57	5.18	.6953
Total	123.72	3.17	
1981			
Males			
10-14	27.84	2.83	-.5386
15-19	240.12	20.31	.6376
20-24	376.19	32.04	.7119
25-29	285.56	26.33	-.5659
Total	929.71	21.01	
Females			
10-14	6.16	0.66	1.1452
15-19	52.88	4.67	-1.3587
20-24	82.81	7.08	-1.5174
25-29	65.44	5.99	1.1822
Total	207.44	4.79	

TABLE 4

Log-Rate Equation for Model 4

Effect	λ	λ/SE_λ	$e^{-\lambda}$
INTERCEPT*	-9.9684	---	.0000469
Age			
10-14	-1.6947	-15.00	.1837
15-19	.3892	7.23	1.4758
20-24	.7921	15.96	2.2080
25-29	.5134	---	1.6709
Sex			
Male	.7392	24.72	2.0943
Female	-.7392	-24.72	.4775
Year			
1971	-.2235	-5.37	.7997
1981	.2235	5.37	1.2504
Age x Year			
1971			
10-14	-.2146	-1.90	.8069
15-19	-.0078	-.14	.9922
20-24	.0713	1.44	1.0740
25-29	.2213	---	1.2477
1981			
10-14	.2146	1.90	1.2394
15-19	.0078	.14	1.0079
20-24	-.0713	-1.44	.9312
25-29	-.2213	---	.8015
L^2	9.87		
DF	7		
P	.196		

--- Standard error not computed.

* The intercept is not given in the SPSSx output and must be calculated manually as follows:

$$\alpha = \log(d_{ij}^{*}/E_{ij} - \lambda^{A}_{i(1)} + \lambda^{S}_{j(1)} + \lambda^{Y}_{k(1)} + \lambda^{AY}_{ik(1)}).$$

where dij= expected suicides for males age 10-14 in 1971, Eij is the exposures for this group, and the $\lambda_{(1)}$'s are the corresponding parameters for this group.

the lowest risk is for those aged 10-14. Thus the chance of suicide is not uniform across age groups. The likelihood of self-inflicted death is twice as great among young men than among young women. The risk of suicide has increased over time by 25 percent. From the interaction terms we note that the odds of committing suicide increase with age in 1971, but decline in 1981.

5. Conclusions

In this paper we have shown how log-rate models, appropriate for dichotomous dependent variables, can be used to model and standardize simultaneously any dichotomous dependent variable that can be treated as a rate. This exposition has not elaborated on the formal properties of the log-rate model, such as the assumed underlying distribution of the event modelled. The assumptions of the log-rate model used in the empirical example assumes a Poisson distribution for suicide (see Clogg and Eliason 1987:28; Hoem, 1987; Laird and Olivier, 1981 for a more complete discussion of this). This distribution is particularly relevant if one is interested in testing statistical hypotheses about rare events.

APPENDIX A: SPSSx Log-Linear Program Used to Model Youth Suicide Rates

---

```
TITLE LOG-RATE STANDARDIZATION: YOUTH SUICIDE IN CANADA.

Data List  List Records = 1

       \Y,S,A,SUI,POP,P1971

COMPUTE TSPOP=7898610

IF (S=1) SPOP = 3998370

IF (S=2) SPOP = 3900240

COMPUTE SW = P1971/SPOP

COMPUTE X = (POP/SW)/TSPOP

WEIGHT by SUI

COMMENT Y = Year, S = Sex, A = Age, SUI = Suicide, POP =
Population, P1971 = 1971 Population as standard, TSPOP = Total
standard population, SPOP = Sex-Specific Totals in Standard
Population.  Variable Levels are: AGE (1=10-14, 2=15-19, 3=20-24,
4=25-29), SEX (1=male, 2=female), YEAR (1=1971, 2=1981).

LOGLINEAR Y(1,2) S(1,2) A(1,4)/
       CWEIGHT = X/
       PRINT=ESTIM/
       DESIGN=A,S,Y,A BY Y


BEGIN DATA
       1      1      1       13       1181450     1181450
       1      1      2      136       1074435     1074435
       1      1      3      218        941775      941775
       1      1      4      178        800710      800710
       1      2      1        4       1129285     1129285
       1      2      2       32       1039915     1039915
       1      2      3       54        947630      947630
       1      2      4       45        783410      783410
       2      1      1       25        984740     1181450
       2      1      2      250       1182015     1074435
       2      1      3      390       1174295      941775
       2      1      4      276       1084410      800710
       2      2      1        9        936130     1129285
       2      2      2       43       1132875     1039915
       2      2      3       69       1169520      947630
       2      2      4       75       1093200      783410
END DATA
```

REFERENCES

Clogg, C.C. and S.R. Eliason. 1987. "Some common problems in log-linear analysis." Sociological Methods and Research 16(1):8-44

Clogg, C.C. and Scott R. Eliason. 1988. A flexible procedure for adjusting rates and proportions, including statistical methods for group comparisons." American Sociological Review 53:267-283.

Haberman, S.J. 1978. Analysis of Qualitative Data. Vol. 1. New York: Academic Press.

Haberman, S.J. 1979. Analysis of Qualitative Data. Vol. 2. New York: Academic Press.

Hoem, Jan M. 1987. Statistical analysis of a multiplicative model and its application to the standardization of vital rates: A review." International Statistics Review 55(2):119-152.

Kitagawa, E.M. 1955. "Components of a difference between two rates." Journal of the American Statistical Association 60:1168-1174.

Kitagawa, E.M. 1964. "Standardized comparisons in population research." Demography 1:296-315.

Laird, M. and D. Olivier. 1981. "Covariance analysis of censored survival data using log-linear analysis techniques." Journal of the American Statistical Association 76(374):231-240.

Lalu, N.M. 1987. "Standardization in criminology." Criminometrica 3:27-32.

Shryock, H.S. and J.S. Siegel. 1973. The Methods and Materials of Demography. Washington, D.C.: U.S. Department of Commerce.

Statistics Canada. 1974. Vital Statistics, 1971. Ottawa.

Statistics Canada. 1983. Vital Statistics, 1981. Ottawa.

Statistics Canada. Census of Canada, 1971 and 1981.

ACKNOWLEDGEMENTS:

The author is grateful to Clifford C. Clogg and Ed Shihadeh for their useful comments on an earlier draft of this paper. Any errors or omissions are the responsibility of the author.

RESEARCH DISCUSSION PAPER SERIES

54. Contraceptive Use in Africa.
David Joe Achanfuo–Yeboah. October 1988.

55. Working Life and Unemployment Tables for Males and Females Canada, 1981.
Pam Smith. October 1988.

56. Beyond Rational Action: Weberian Theory and Policy Issues in AIDS.
Soma Hewa and Robert W. Hetherington. January 1989.

57. A Causal Modelling Approach of Analyzing Statistical Interaction: Ethnic Differentials in the Timing of Early Family Formation in Sri Lanka and Malaysia.
Wei–Jun Jean Yeung. April 1989.

58. The Effects of Religious Factors on Childlessness: The Canadian Case.
Vijaya Krishnan. April 1989.

59. A Study of the Contemporary Population Migration in the People's Republic of China.
Yihua Yuan. May 1989.

60. Might Suicide Underreporting Affect Hypothesis Test? The Case of Suicide and Divorce.
Mark Speechley and Frank Trovato. July 1989.

61. Demographic Concerns in Belles–Lettres: Insights From the Writings of Gûnter Grass and Margaret Atwood.
Karol Krótki and Guy Germain. January 1990.

62. Some Lacunae in Demographic Data Collection in South Asia with Special Reference to India.
Parameswara Krishnan. January 1990.

63. The Emergence of Multiethnicities in 1981 and 1986: Their Sociocultural Significance.
Karol Krótki and Dave Odynak. February 1990.

64. Dynamics of Internal Migration in Bangladesh.
A.K.M. Nurun Nabi. April 1990.

65. Why Are Canadians Dying Out?
Karol Krótki. April 1990.

66. A Theoretical Model for the Study of Migration and Urbanisation
David Achanfuo–Yeboah. May 1990.

67. Political Demography of the Emergence of Bangladesh
A.K.M. Nurun Nabi and P. Krishnan. May 1990.

68. Dependency Ratios: Conventional and Real.
Parameswara Krishnan and Manoshi Das. June 1990.